Accepted to Phys. Fluids 10.1063/5.0188610

Unsupervised Machine Learning of Virus Dispersion Indoors

Nicholas Christakis (Νικόλαος Χρηστάκης),[1,2,a] Dimitris Drikakis (Δημήτρης Δρικάκης),[1,b] Konstantinos Ritos (Κωνσταντίνος Ρήτος),[3,1,c] and Ioannis W. Kokkinakis (Ιωάννης Κοκκινάκης)[1,d]
[1)]*Institute for Advanced Modeling and Simulation, University of Nicosia, Nicosia, CY-2417, Cyprus*
[2)]*Laboratory of Applied Mathematics, University of Crete, Heraklion GR-70013, Greece*
[3)]*Department of Mechanical Engineering, University of Thessaly, Volos 38221, Greece*

This paper concerns analyses of virus droplet dynamics resulting from coughing events within a confined environment using, as an example, a typical cruiser's cabin. It is of paramount importance to be able to comprehend and predict droplet dispersion patterns within enclosed spaces under varying conditions. Numerical simulations are expensive and difficult to perform in real-time situations. Unsupervised Machine Learning methods are proposed to study droplet dispersion patterns. Data from multi-phase computational fluid dynamics simulations of coughing events at different flow rates are utilized with an unsupervised learning algorithm to identify prevailing trends based on the distance traveled by the droplets and their sizes. The algorithm determines optimal clustering by introducing novel metrics such as the Clustering Dominance Index and Uncertainty. Our analysis revealed the existence of three distinct stages for droplet dispersion during a coughing event, irrespective of the underlying flow rates. An initial stage where all droplets disperse homogeneously. An intermediate stage where larger droplets overtake the smaller ones, and a final stage where the smaller droplets overtake the larger ones. This is the first time Computational Fluid Dynamics is coupled with Unsupervised Learning to study particles' dispersion and understand their dynamic behaviour.

I. INTRODUCTION

Studying the transmission of viruses is paramount in the post-COVID-19 pandemic era, as safety and prevention have emerged as critical concerns. Detecting the earliest signs of virus carriers and investigating the dynamics of virus dissemination during actions such as coughing and the dispersion of saliva droplets in enclosed environments are pivotal for public health. Mathematical models have historically been vital in addressing these issues, but their computational demands often result in computationally expensive simulations.[1,2] Additionally, experiments conducted to validate these models also encompass challenges and uncertainties. Limiting factors in experiments for COVID-19 research may include the irreproducibility of in-vitro studies in in-vivo studies due to the complexity of physiological processes and the behavior of genetically modified tissues..[3]

Machine Learning (ML) can unveil hidden patterns within noisy or incomplete data.[4–7] Unsupervised Learning (UL), in particular, has achieved remarkable success across diverse scientific domains and is better suited for problems of higher complexity in identifying groups within data sets. These problems include deciphering complex genetic structures[8], understanding natural language semantics[9], and even discovering astronomical phenomena[10], among many others. UL is promising because it can handle unlabelled data, including experimental, computational or field measurements. Thus, it does not require user intervention. Given the persistent gaps in our understanding of the physics underlying various processes, such as virus transmission, the above becomes especially pertinent. Several UL algorithms have been developed.[11–13]. However, UL could encompass higher uncertainty in engineering, physics, and biomedical applications where the underlying physical processes remain elusive.[13] The application of UL in physics and mechanics-driven processes differs from the application of general data sets such as in economics, marketing, or media. Classifying patterns and predictions in fluid mechanics and engineering must be physics-relevant and explainable. Verification and validation of ML can be based on first principles and experimental data (if available). Simplicity in the developed ML algorithms is also important, as it will facilitate explainability and broader ML application with confidence.

Recent developments in UL, such as the Reduce UNcertainty and Increase CONfidence (RUN-ICON) algorithm,[12] have further expanded UL's applicability. This state-of-the-art algorithm has demonstrated its prowess in reliably clustering artificially generated and real-world data. Applying the RUN-ICON algorithm on physics-based data produced by Computational Fluid Dynamics (CFD) can give invaluable insights into the behaviour of saliva droplets emitted from coughing or sneezing. This analysis provides a data-driven perspective on virus transmission dynamics that was previously unattainable. To our knowledge, combining UL and CFD in particle-driven dynamics has yet to be carried out. It offers a promising avenue for advancing our understanding in an era where precision and rapid response to emerging infectious diseases are paramount. The above motivates the present work.

The paper is organized as follows. Section II sets the frame by discussing COVID-19 spreading and infection in enclosed spaces. Section III presents the CFD methods and models used in the study. Section IV presents the Unsupervised Learning methods employed. The results are presented in Section V. The conclusions drawn from this study are summarized in Section VI.

[a)]Electronic mail: nchristakis@tem.uoc.gr
[b)]Electronic mail: Author to whom correspondence should be addressed: drikakis.d@unic.ac.cy.
[c)]Electronic mail: konritos@uth.gr
[d)]Electronic mail: kokkinakis.i@unic.ac.cy

II. BACKGROUND AND MOTIVATION

The recent COVID-19 pandemic with more than 770×10^6 confirmed cases and nearly 7×10^6 deaths[14] led the scientific community to a rigorous and urgent study of how droplets and nuclei expelled by infected individuals spread, potentially infecting others, and also produce appropriate mitigation guidelines and strategies.

This pandemic prompted the investigation of virus transmission in closed spaces.[15–22], These studies discussed the air filtration, air purification, and efficiency of HEPA filters, which are supposed to capture more than 99% of particles, provide high ventilation rates with 50% fresh air and 50% filtered while delivering more than 20 changes of air in the cabin.[23] The latter is also referred to in the literature as Air Changes per Hour (ACH) and indicates how many times the air contained in a room could be theoretically completely renewed by fresh or filtered air in one hour. This calculation is based on the ventilation system's volumetric flow rate over the room's volume. It is only an indicative value as flow recirculation, room geometry, ventilation outlets, inlet locations, furniture, and people can highly affect the actual time it takes for complete air renewal. The above studies also showcased the inevitability of virus transmission when two persons are close.

Similarly, aerosol transmission and ventilation configuration in car cabins[24,25] and buses[26,27] has also been the topic of recent research. On the other hand, cruise ships and the transmission inside passenger cabins have been presented only with mechanistic modeling[28] and data analysis from outbreaks.[29] A recent review on the transmission of COVID-19 in cruise ships[30] indicates that cabins of high occupancy have an increased transmission risk, without commenting on the cabin's ventilation system and how this could have affected the transmission.

CFD studies with ventilation suggestions have been limited to small vessels until now.[31] Moreover, there have been contradicting arguments in the literature regarding the recirculation of possibly contaminated air and the ventilation rate efficiency in cruiser ships. Azimi et al.[28] suggested high ventilation rates in cruise ship cabins, but views differ among authors.[32–35]

The most recent standards and regulations on room safety regarding the airborne transmission of viruses focus on high rates of air exchanges.[36–38] This can be inefficient as large energy consumption is needed to maintain high air flow rates, while comfort can be reduced due to the creation of strong air drafts.

The ASHRAE Standard[38,39] provides a formula for the minimum ventilation rate based on occupants and the surface area of a hotel bedroom $\dot{Q}_{outdoor-air} = R_p \times N_p + R_a \times A$, where $\dot{Q}_{outdoor-air}$ is the specified outdoor air that should be supplied in the room, R_p is the required outdoor flow rate per person, R_a is the required outdoor flow rate per unit area and A is the floor area of the room. The appropriate values for $R_p = 2.5\ \ell/(s\cdot \text{person})$ and $R_a = 0.3\ \ell/(s\cdot m^2)$ are defined in the Standard[38,39] and lead to a value of 27 m^3/h for the cabin presented here, assuming 2 occupants. The cruiser designers have proposed a stricter ventilation limit of 30 m^3/h per room occupant, while the World Health Organisation (WHO)[40], the European Federation of HVAC Associations[41], and research studies[42] recommend 36 m^3/h per person. The Federal Public Service (FPS) Health, Food Chain Safety and Environment of Belgium defined the minimum flow rate for their Standard A Level at a slightly higher value of 40 m^3/h per person.[43]

Considering the above discussion and two people occupying a small cruiser cabin, our reference flow rate value is 60 m^3/h. The most recent CDC guidelines based on the draft of ASHRAE Standard 241-2023[38], similarly by other studies[44], propose a minimum of 5 Air Changes per Hour (ACH), which translates to a flow rate of 200 m^3/h for the cabin we investigate in this paper. The ASHRAE Standard 241-2023[38] recommends 15 l/s per occupant, which is equal to 108 m^3/h for the studied cabin. In comparison, a typical home has less than 0.5 ACH based on CDC data, while this number reduces to 0.35 ACH based on the recommendation from ASHRAE Standard (62.2-2019).[45]

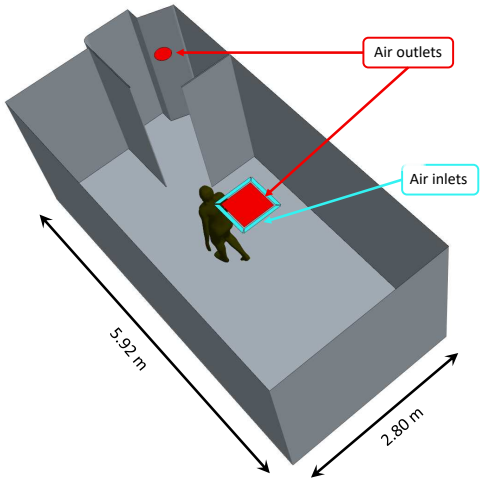


FIG. 1. The simulation domain used in the CFD simulations with the main dimensions. The height of the room is 2.4 m. The location of the coughing person is also shown for Case B, as well as the ventilation inlets and outlets.

A typical cruise ship cabin was considered, and the details can be found in[46]. In addition, the effect of droplet evaporation has been considered, as review papers suggest,[47] and not incorrectly omitted. The role of droplets in virus spreading and the importance of evaporation and subsequent size alternation have been highlighted in the publication by Dhand and Li.[48] In our simulations, we have utilized 1.5 to 15 ACH to capture all possible cases from poor/minimal ventilation up to exceeding the most recent recommendations. Recently,[46] it has been demonstrated that droplets can spread up to 5 times more when high ventilation rates are utilised in the initial few seconds after a person has coughed. CFD simulations of

coughing events for varying flow rates were performed, where several droplets of varying sizes were emitted, and their trajectories were recorded. The data sets from the above study are utilised here to produce, with the aid of the RUN-ICON algorithm, a classification of patterns and assess the behaviour of the droplets.

III. CFD OF DROPLETS SPREADING

We have employed CFD simulations to generate appropriate data on saliva droplets spreading from a coughing person in a closed space. In more detail, we simulate the cabin from a cruise ship with the overall floor dimensions shown in Fig. 1. The cabin height is 2.4 m, with the overall deck height of the ship being 2.8 m. A representative air conditioning unit is placed at the cabin's centre with a square outlet (□48 cm) and four rectangular inlets (55 $cm \times 5$ cm each) that expel air at an angle of 45^o. The bathroom area has an additional circular outlet (⌀25 cm). In Fig. 1, the person is at the centre of the room, which is one of the representative cases studied here. The geometry has been meshed with polyhedral non-uniform cells ($\approx 0.6 \times 10^6$), with significant refinement in all inlet and outlet regions, where a conical refinement area is defined in each case. For example, at the mouth and up to a distance of 0.5 m in the streamwise direction, the cells have a maximum isotropic size of 4 mm compared to the overall targeted cell size of 4 cm.

We employ a laminar multi-component gas model representing two distinguished Eulerian fluids: dry air and water vapor (humidity), and saliva droplets as liquid water Lagrangian particles. The latter is only inserted from the person's mouth during coughing. Although saliva is a complex fluid and varies from person to person, taking its viscosity close to that of water is a valid approximation.[49]

Further, we utilize a compressible, unsteady multiphase solver with the Ranz-Marshall model applied for the Nusselt and Sherwood numbers.[50,51] The ideal gas law expresses the density as a function of temperature and pressure in our calculations. At the same time, the dynamic viscosity of air and water vapor is calculated based on Sutherland's law. All simulations have been performed in Star-CCM+ 2210[52] with a 2^{nd} order implicit temporal solver.

No-slip adiabatic boundary conditions have been applied on all walls, the ceiling, the floor and the person's body. The maximum timestep used in the implicit temporal solver was 0.01 s, which was reduced to 0.5 ms during the coughing event. Outlet boundary conditions with a specified mass flow rate have been applied in the Air outlets, shown in Fig. 1, with $1/3$ of the flow directed to the bathroom outlet and the rest to the A/C outlet leading to a pressure balance in the cabin. The ventilation inlets expel air with the conditions shown in Table I, providing the overall targeted mass flow rate with the air blowing at an angle of 45^o.

During the initial 60s simulation, no saliva droplets are in the domain to produce a fully developed airflow. The human cough is imitated over 0.12s. The velocity applied at the mouth during the cough is $u_x = 8.5$m/s, as measured by Scharfman et al.[53] After the coughing event, and during the initial 60s, the air is expelled from the mouth with a low breathing velocity of $V_{breath} = 0.1$m/s. The initial total mass of the injected saliva into the domain is 7.7mg, which agrees well with existing experimental measurements reported in the literature.[54,55] The simulation continues for another 20 s, at least, leading to a total simulation time of 80.12s, where more than 99.99% of the saliva droplets have evaporated in all cases presented here.

T_∞ (C^o)	P_∞ (atm)	RH (%)
20	1	55
T_{mouth} (C^o)	V_{cough} (m/s)	V_{breath} (m/s)
34	8.5	0.1

TABLE I. Simulation conditions. The person's mouth is always considered an inlet with the same ambient air conditions as the ventilation inlets.

We employ the Weibull distribution[56] on the size of the injected saliva droplets

$$f = \frac{n}{\bar{d}_p}\left(\frac{d_p}{\bar{d}_p}\right)^{n-1} e^{-(d_p/\bar{d}_p)^n}, \quad (1)$$

where d_p is the saliva droplet diameter, $n = 8$ and $\bar{d}_p = 80\,\mu m$. Our choice is based on several previous studies,[18,55,57] indicating the appropriateness of this distribution for dispersing water-like cloud droplets. We should also mention that the Lagrangian phase equations were discretized by employing implicit numerical schemes in the second order with two-way coupling and a quasi-steady evaporation model. Further geometrical, meshing and solver details are given in the recent publication by Ritos et al.[46]

In this study, we selected to simulate three different flow rates for the air from the ventilation system and then utilize the RUN-ICON Unsupervised Learning algorithm to examine the dispersion of saliva droplets. In addition, two positions of the coughing person, one at the far end of the room (Case A) and one in the middle (Case B), are studied. The incoming air is considered clean, i.e., free from contaminants, without specifying whether it is outdoor air or treated with filtered recirculated air.

Representative results from the multiphase CFD simulations are given in Fig. 2, where the flow field is shown with a grey-scale contour plot at different heights from the floor, while saliva droplets are also visualized. An extreme, high ventilation case ($600\ m^3/h$) has been selected to highlight the quick spreading of saliva droplets after only 8 s from the coughing event. A significant amount of droplets has entered the bathroom area from the door opening, covering a distance greater than 2 m from the coughing person. Most droplets have already settled on the floor, while some maintain a considerable elevation of 40 cm from the floor level.

During the first two seconds of the coughing event, the saliva droplets have penetrated the air up to a distance of 80 cm while many maintain a high elevation of over 1 m from the floor. By the latest time shown here (8 s), saliva droplets have

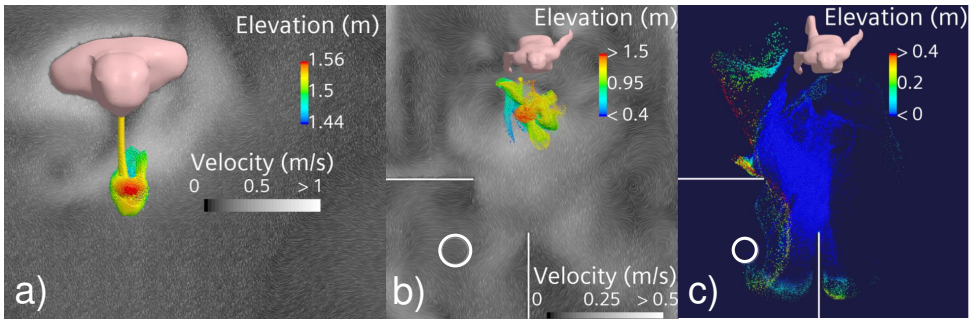


FIG. 2. Characteristic snapshots from the reference CFD simulations[46] highlight the spreading of the saliva droplets for Case B600 at various instances. The droplets are colored based on their distance from the floor (elevation). As expected, at the initial coughing stage, the droplets are close to the height of the person's mouth (1.53 m) and quickly fall to the ground due to gravity after a few seconds. a) The contour plot shows the velocity flow at the height of $z = 1.4\,m$ from the floor, while $t = 0.12\,s$. b) The contour plot shows the velocity field at the height of $z = 0.5\,m$ from the floor, while $t = 2\,s$. c) Droplet spreading at $t = 8\,s$.

lost around 90% of their initial mass. As a result, later stages are not shown as most droplets, if not all, will have settled on the floor, and their overall mass will have significantly reduced. Further CFD results and the corresponding analysis can be found in[46].

In the following sections, the CFD results will be analysed with the RUN-ICON algorithm to gain insight into the dynamics of droplet dispersion based on their sizes and the distances they travel from the origin. We now move on to give a brief description of the algorithm.

IV. UNSUPERVISED LEARNING METHOD

UL represents a fundamental machine learning paradigm that has persistently endured over the years and continues to hold paramount significance to this date in diverse domains of research and applications, such as image processing[58,59], sleep stages classification[60] and mechanical damage detection[61].UL algorithms, however, do face several challenges when it comes to particle clustering tasks, such as sensitivity to hyperparameters, dimensionality and feature selection, and scalability, among others[13] The RUN-ICON (Reduced UNcertainty-Increased CONfidence) algorithm,[12] which is utilized in this work, aims to address many of these challenges by ensuring the selection of the most dominant clusters occurs with high confidence and low uncertainty. The main advantage of the RUN-ICON, compared to other UL techniques, is that it alleviates the requirement to decide on the optimum number of clusters based only on intuitive criteria. RUN-ICON offers a systematic approach to determining the optimum number of clusters, reducing reliance on subjective judgments. The algorithm was constructed to effectively determine the optimal number of clusters by identifying commonly dominant centres across multiple repetitions of the K-means++ algorithm. The algorithm does not rely on the Sum of Squared Errors and determines optimal clustering by introducing novel metrics such as the Clustering Dominance Index (CDI) and Uncertainty. CDI is linked with the frequency of occurrence of a specific clustering configuration when requesting the splitting of the data set in a certain number of clusters and could be translated as the probability of that specific configuration occurring. Uncertainty is the relative difference between upper and lower CDI bounds for a clustering configuration and represents the maximum variance from the mean for that specific configuration. The algorithm, thus, concentrates on discerning the most resilient clusters that signify the true inherent structure of the data. By giving precedence to stability, the algorithm guarantees that the chosen cluster centres accurately represent the underlying patterns while mitigating the influence of outliers or noisy data points. This enables the generation of more robust and interpretable clustering results. The algorithm involves the following tasks:

1. Perform the K-means++ clustering algorithm 100 times, each time specifying a particular number of clusters.
2. After each run, compute the coordinates of the cluster centres.
3. Conduct a comparison of the cluster centre coordinates from all 100 runs to identify the coordinates that appear most frequently.
4. Select the cluster centers that exhibit the highest frequency of appearance as the dominant centers corresponding to the chosen number of clusters and calculate CDI.
5. Replicate the preceding steps (steps 1-4) ten more times to obtain ten sets of dominant cluster centres.
6. Compute the average frequency of appearance for the dominant centers and their respective CDIs, derived from the ten repetitions, ensuring that they correspond

to the same clustering centers. Calculate these metrics' upper and lower bounds and assess the variance within this range.

7. Repeat steps 1-6 for various cluster numbers, beginning with 3 clusters and extending up to 10 clusters.

8. Determine the cluster number with the highest average CDI across all possible clustering scenarios, ranging from 3 to 10 clusters. Identify the corresponding dominant cluster centers and assess the variance between the upper and lower bounds, aiming for a variance of less than 30

9. Choose the cluster number with the highest average CDI and low variance as the optimal number of clusters for the RUN-ICON algorithm.

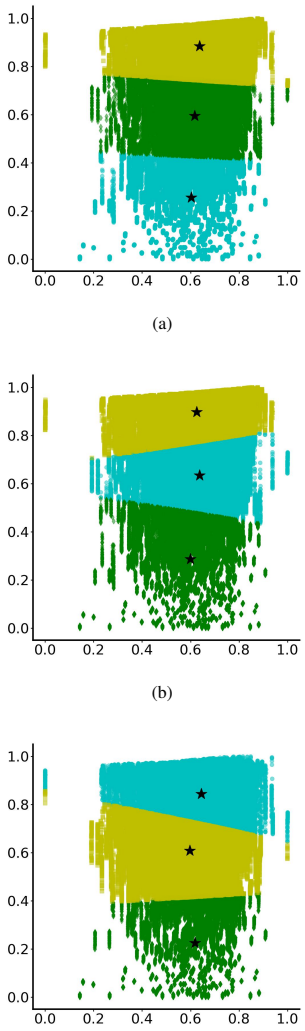
RUN-ICON algorithm's advantages over different UL algorithms were presented in a previous study.[12] The reliability and interpretability were established for different computational data sets where the intended clustering configuration was known a priori. The algorithm's accuracy in identifying the intended clustering configurations is always about 97% compared to 80% provided by other algorithms. Moreover, RUN-ICON was applied to mathematically defined particle-like dispersion[13], with particles randomly left to disperse in space. RUN-ICON recognized dominant clustering patterns compared to other UL techniques, whereas other UL algorithms failed. This further motivates the application of RUN-ICON to different realistic flow scenarios and applications.

V. UNSUPERVISED LEARNING IMPLEMENTATION AND DROPLET CLASSIFICATION

To apply the RUN-ICON algorithm to the specific problem of droplet spreading, three different flow rates, four different times from the start of the coughing event and two different positions of the coughing person were considered (as given in Table II).

Ten different cases were examined, and the algorithm was required to find clusters of particles based on the final positions (distance from the origin) of droplets with different diameters. As distance from the origin, the Euclidean norm of a three-dimensional position vector was considered. The number of droplets every time varied between 114,000 and 146,000, with particles having a wide range of diameters (from 108 μm to 0.3 μm). All data were normalized between 0 and 1 to make all data scale-invariant since the problem involves parameters of varying orders of magnitude. The algorithm was run on an 8-core Intel i5-9300HF 8 GB RAM processor, and its performance varied between 6.5 and 15 minutes, depending on the number of droplets and the number of clusters required every time. This indicated the algorithm's ability to handle large data sets efficiently.

Table II presents the optimum number of clusters for each time, the corresponding flow rate and the person's position. RUN-ICON predicted these clustering configurations with

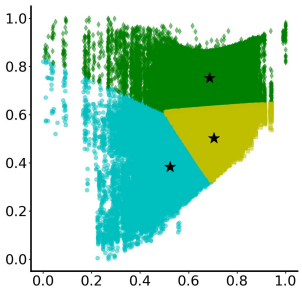


(a)

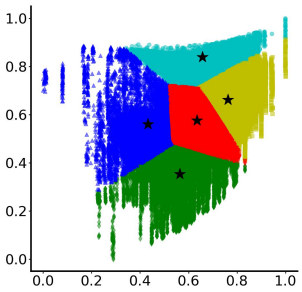
(b)

(c)

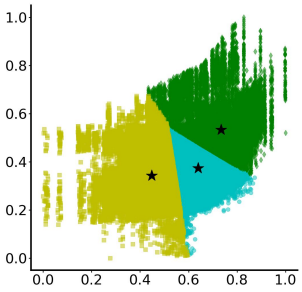
FIG. 3. Clustering of droplets at time 0.12 s from the start of the coughing event at three different flow rates and two different locations of the coughing person (a) A60, (b) B120 and (c) B600. In (a) max-min of x-axis: 108-19 μm, max-min of y-axis: 0.35-0.004 m. In (b) max-min of x-axis: 108-19 μm, max-min of y-axis: 0.37-0.004 m. In (c) max-min of x-axis: 108-19 μm, max-min of y-axis: 0.32-0.004 m. The stars represent the cluster centres, and each colour indicates droplets belonging to the same cluster.



(a)



(b)



(c)

FIG. 4. Clustering of droplets at time 2 s from the start of the coughing event at two different flow rates and two different locations of the coughing person (a) A120, (b) B120 and (c) B600. In (a) max-min of x-axis: 102-0.6 μm, max-min of y-axis: 0.83-0.15 m. In (b) max-min of x-axis: 102-1 μm, max-min of y-axis: 1.1-0.15 m. In (c) max-min of x-axis: 102-3 μm, max-min of y-axis: 1.35-0.14 m. The stars represent the cluster centres, and each colour indicates droplets belonging to the same cluster.

confidence very close to or equal to 100%. For every examined case, at least one other clustering configuration had equally high confidence and predicted separation in more clusters. However, we have decided to choose the separation in fewer clusters; choosing the simplest state of a system without specific scientific or problem-driven reasons is a well-established practice in science due to its empirical success, methodological soundness, and mathematical tractability.[62] Simplicity is often a powerful guiding principle in our attempt to understand the physical world. Moreover, it was observed that no new centre configuration was predicted for the higher number of clusters. Rather, every new cluster resulted from the breakup of an old cluster, not adding any further physical information to the clustering.

Choosing "simplicity over complexity" is derived from Occam's razor principle,[63], which dictates that the simplest one is preferred among competing hypotheses or models that adequately explain observed phenomena. In other words, one should choose the explanation or model that makes the fewest assumptions or introduces the fewest entities as long as it remains consistent with the observed data. This idea has recently been explored in statistical learning and data science.[64,65] In clustering, simplicity is often associated with fewer clusters and straightforward patterns. This perfectly aligns with Occam's razor principle. When we consider a simple clustering with fewer clusters, the probability distribution over the data points tends to have a more straightforward structure. The likelihood of observing the data given this simpler clustering is often higher because it requires fewer parameters to describe. On the contrary, a more complex clustering involving more clusters introduces additional parameters to the model. This increased complexity may lead to overfitting, where the model starts capturing noise or incidental structures in the data rather than genuine underlying patterns. As a result, the likelihood of observing the data given this complex clustering becomes lower. This reasoning may be quantified using a probabilistic framework.

Consider a data set X with N data points, and let Z be a random variable representing the cluster assignments. The goal is to find the clustering Z that maximizes the likelihood of the data set X. Mathematically, this is expressed as $P(X|Z)$.

This is the likelihood of observing the data X given the cluster assignment Z. To express within a probabilistic framework the procedure of selecting a preferred clustering configuration, the Bayesian Information Criterion (BIC)[66] will be utilized, which penalizes the likelihood based on the number of parameters: $BIC = -2\ln P(X|Z) + k\ln N$, where k is the number of parameters and N is the number of data points. The preferred model is the one which exhibits the minimum BIC. The penalty term $k\log N$ discourages overly complex models, reinforcing the preference for simplicity. Then, if the likelihood for two clustering configurations, namely Z_1 and Z_2 (a simpler clustering and a more complex one, respectively), is the same, minimising BIC depends solely on k. And since a more complex clustering for the same data points N will involve more parameters, i.e., greater number of clusters, than a simpler one, the penalty term will become greater in the case of the more complex clustering. In this way, $BIC_1 < BIC_2$

Time (s)	Flow rate (m^3/h)	Case	Optimum # of clusters
0.12	60	A	3
0.12	120	B	3
0.12	600	B	3
2	120	A	3
2	120	B	5
2	600	B	3
4	120	B	3
4	600	B	3
8	120	B	3
8	600	B	3

TABLE II. Times after the start of the coughing event, flow rates, case (indicating the person's position) and optimum number of clusters.

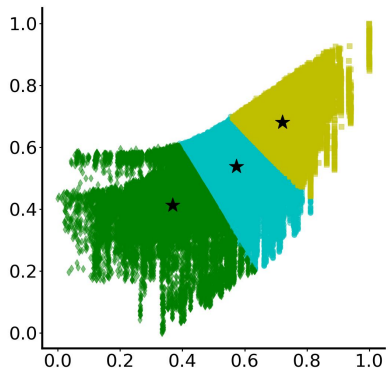
implies that the simpler model is more likely to constitute a more faithful representation of the underlying structure in the data set. This inequality expresses the probabilistic rationale behind preferring simpler clustering.

Figure 3 presents the clustering of droplets at time 0.12 s from the start of the coughing event. In this figure and all subsequent ones, the x-axis represents the normalized droplet diameter, and the y-axis represents the normalized distance from the origin of the droplets. Three distinct clusters for all 3 cases were predicted, with a stratification of droplets based on the distances they have traveled from the origin. Since only a very short time from the event had passed, the droplets seemed to disperse across space homogeneously, irrespective of their size. Only a few of the smallest droplets seemed to have moved further away from the origin.

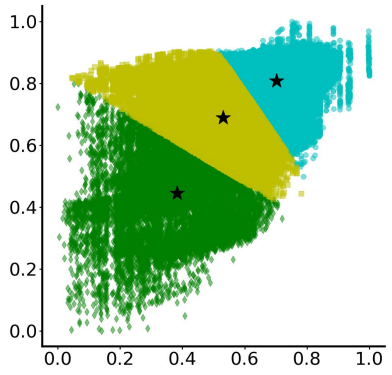
Figure 4 presents the clustering of droplets at time 2 s from the start of the coughing event. As can now be observed, 3 (for flow rates A120 and B600) or 5 (for flow rate B120) clusters were predicted. As the flow progressed, the largest particles, due to the flow dynamics, started moving faster than the smaller ones, thus overtaking them and moving further away from the origin. As observed, flow rate B120 5 clusters are predicted, rather than 3, as is the case for all other flow rates and times. This is because a transitional state in the system occurs at this specific flow rate and time, where multiple configurations or phenomena coexist. This could manifest as additional clusters in the feature space.

In Figure 5, the clustering of droplets at time 4 s from the start of the coughing event is presented for two flow rates (namely, B120 and B600). As can now be observed, the algorithm predicted separation into 3 clusters. At this flow instance, the smaller particles seemed to increase their momentum and, thus, start to catch up with the largest particles, especially with the higher flow rate.

Finally, in Figure 6, the clustering of droplets at time 8 s from the start of the coughing event is presented for two flow rates (namely, B120 and B600). The algorithm predicted separation into 3 clusters. At this flow instance, the smaller particles seemed to have fully caught up with the largest particles. The smaller particles have overtaken the larger ones for the highest flow rate, and those particles have moved further away from the origin with the airstream. The larger droplets have



(a)



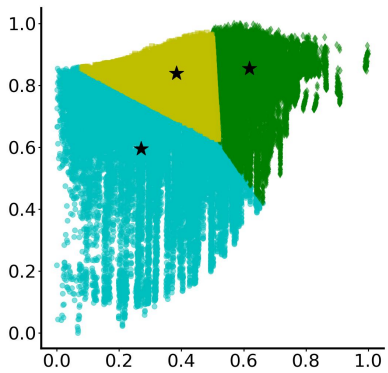
(b)

FIG. 5. Clustering of droplets at time 4 s from the start of the coughing event and location B at two different flow rates (a) 120 m^3/h and (b) 600 m^3/h. In (a) max-min of x-axis: 95-0.3 μm, max-min of y-axis: 1.63-0.32 m. In (b) max-min of x-axis: 95-1.14 μm, max-min of y-axis: 1.92-0.27 m. The stars represent the cluster centres, and each colour indicates droplets belonging to the same cluster.

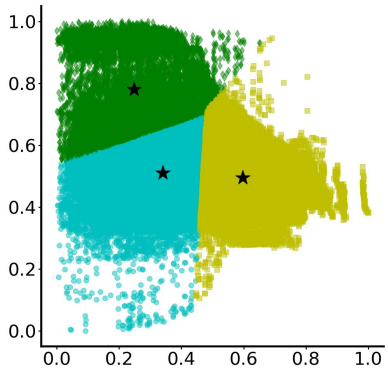
fallen to the ground by that time.

The results suggest a discernible pattern in droplet dispersion dynamics during coughing events, demonstrating a remarkable consistency across varying flow rates and droplet diameters. These results underscore the presence of three distinct stages within the temporal evolution of droplet dispersion, irrespective of the initial conditions.

In the initial phase, there is an evident homogeneity in droplet dispersion, characterized by a uniform dispersion of



(a)



(b)

FIG. 6. Clustering of droplets at time 8 s from the start of the coughing event and location B at two different flow rates (a) 120 m^3/h and (b) 600 m^3/h. In (a) max-min of x-axis: 82-0.6 μm, max-min of y-axis: 1.89-0.64 m. In (b) max-min of x-axis: 81-0.53 μm, max-min of y-axis: 2.62-1 m. The stars represent the cluster centres, and each colour indicates droplets belonging to the same cluster.

droplets, regardless of their sizes. An initial even distribution of droplets in the surrounding environment marks this phase. This homogeneous dispersion is considered to be indicative of the immediate effects of the force exerted during the coughing event.

Following the initial stage, a transitional phase emerges, wherein larger droplets exhibit greater momentum, surpassing their smaller counterparts and manifesting an overtaking phenomenon. This enhanced momentum of larger droplets during this intermediary stage suggests a size-dependent influence on their trajectories, leading to an asymmetry in the spatial distribution of droplets. The discernible divergence in trajectories based on droplet size is a critical observation, shedding light on droplet behavior dynamics during the mid-phase dispersion.

Ultimately, the system undergoes a third stage, characterized by a noteworthy reversal of the observed size-dependent droplet behaviour disparities. In this final stage, the larger droplets, having initially outpaced their smaller counterparts, exhibit a deceleration in their momentum. This phenomenon results in a convergence of velocities between larger and smaller droplets, eventually culminating in the smaller particles overtaking their larger counterparts.

This droplet's behaviour seems to be confirmed by recent experimental data from a hospital ward.[67] even though the flow rates and droplet sizes differ, the experimental data seem to indicate that during the initial stages of coughing, big and small droplets were concentrated at similar lateral distances from the origin. Then, as time progressed, the large droplets seemed to overtake the small ones marginally. Then, after approximately 15 s, the small droplets overtake the larger ones in the lateral direction. The presence of the largest droplets in the experiments stopped after approximately 30 seconds since the flow conditions were probably such that they had all fallen on the floor, and they were not taken into account in the flow patterns any more.

The observed reversal of behavior implies a dynamic interplay of forces within the ambient environment, potentially influenced by inter-droplet interactions or other underlying physical mechanisms. The complexity inherent in these dynamics emphasizes the necessity for additional research to clarify the exact contributing factors and mechanisms that govern the observed stages of droplet dispersion during coughing events.

VI. CONCLUSIONS

This work used CFD data and Unsupervised Machine Learning to study airflow dynamics in a cruise ship cabin, where coughing events occurred, and the trajectories of emitted droplets of different sizes were recorded. Three different flow rates and two different locations of the coughing person were used, while the final positions of the droplets were recorded at times 0.12 s, 2, s, 4, s and 8 s after the event's start. Then, the model results were used as input to the RUN-ICON UL algorithm. Despite the many particles involved (well over 100,000 droplets emitted during each event), the algorithm managed to predict a dominant clustering configuration with high confidence for all different flow rates. The conclusions of these predictions are summarized as follows:

- At the start of the event (0.12 s), all droplets seemed to disperse homogeneously, irrespective of size.
- As time progressed (at 2 s), a clear separation of larger particles appeared, overtaking the smaller ones and moving further away from the origin.

- When more time was given to the particle flow to evolve (at 4 s), the smaller particles seemed to be catching up with the larger ones in the distances that they traveled.
- At the final recorded time (8s), it was obvious that the smaller particles (especially those in the ventilation case with the highest flow rate) had already overtaken the larger particles and moved further away from the origin. Most of the larger particles had settled on the floor by that time.

According to these findings, irrespective of flow rates and droplet diameters, three distinct stages of droplet dispersion exist during a coughing event. Initially, droplets disperse homogeneously, irrespective of their sizes; then, the larger droplets overtake the smaller ones and move faster away from the origin; in the final stage, the larger particles seem to lose their momentum, thus allowing the smaller particles to catch up and overtake them.

The combination of CFD and UL, as presented in this work, for studying viral droplet dispersion in enclosed spaces is believed to be the first of its kind. This study provides insights into the temporal evolution of droplet dispersion, revealing distinct stages characterized by size-independent homogeneity, size-dependent overtaking, and a subsequent reversal of the size-dependent dynamics. These observations have significant implications for our understanding of respiratory droplet dynamics, with potential applications in public health, epidemiology, and the design of effective mitigation strategies in the context of infectious disease transmission. Further work is already underway, with more physical experiments and numerical simulations being performed for different spaces and airflow configurations to accurately assess the impact of ventilation in the transmission of respiratory diseases in enclosed spaces, such as ship cabins.

VII. ACKNOWLEDGEMENTS

This paper is supported by the European Union's Horizon Europe Research and Innovation Actions programme under grant agreement No 101069937, project name: HS4U (HEALTHY SHIP 4U). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure, and Environment Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

VIII. AUTHOR DECLARATIONS

The authors have no conflicts to disclose.

IX. DATA AVAILABILITY

The authors' data supporting this study's findings are available upon reasonable request.

REFERENCES

[1]M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2008).
[2]J. Wang, "Mathematical models for covid-19: applications, limitations, and potentials," Journal of Public Health and Emergency **4** (2020).
[3]L. A. Caldera-Crespo, M. J. Paidas, S. Roy, C. I. Schulman, N. S. Kenyon, S. Daunert, and A. R. Jayakumar, "Experimental models of covid-19," Frontiers in Cellular and Infection Microbiology **11** (2022), 10.3389/fcimb.2021.792584.
[4]N. Christakis, V. Barbaris, and A. Spentzos, "A new approach in financial modelling with the aid of artificial neural networks," Journal of Algorithms & Computational Technology **5**, 513–529 (2011), https://doi.org/10.1260/1748-3018.5.3.513.
[5]N. Christakis, P. Tirchas, M. Politis, M. Achladianakis, E. Avgenikou, and G. Kossioris, "Covid-liberty, a machine learning computational framework for the study of the covid-19 pandemic in europe. part 2: Setting up the framework with ensemble modeling," International Journal of Neural Networks and Advanced Applications , 27–42 (2021).
[6]K. Poulinakis, D. Drikakis, I. W. Kokkinakis, and S. M. Spottswood, "Machine-learning methods on noisy and sparse data," Mathematics **11** (2023), 10.3390/math11010236.
[7]D. Drikakis and F. Sofos, "Can artificial intelligence accelerate fluid mechanics research?" Fluids **8** (2023), 10.3390/fluids8070212.
[8]X. Shen, C. Jiang, Y. Wen, C. Li, and Q. Lu, "A brief review on deep learning applications in genomic studies," Frontiers in Systems Biology **2** (2022), 10.3389/fsysb.2022.877717.
[9]A. Vlachos, "Evaluating unsupervised learning for natural language processing tasks," in *Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011) pp. 35–42.
[10]K. T. Matchev, K. Matcheva, and A. Roman, "Unsupervised machine learning for exploratory data analysis of exoplanet transmission spectra," The Planetary Science Journal **3**, 205 (2022).
[11]M. Alloghani, D. Al-Jumeily Obe, J. Mustafina, A. Hussain, and A. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," (2020) pp. 3–21.
[12]N. Christakis and D. Drikakis, "Reducing uncertainty and increasing confidence in unsupervised learning," Mathematics **11** (2023), 10.3390/math11143063.
[13]N. Christakis and D. Drikakis, "Unsupervised learning of particles dispersion," Mathematics **11** (2023), 10.3390/math11173637.
[14]WHO, "Who coronavirus disease (covid-19) dashboard data," (2021).
[15]Y. Yan, X. Li, X. Fang, P. Yan, and J. Tu, "Transmission of covid-19 virus by cough-induced particles in an airliner cabin section," Engineering Applications of Computational Fluid Mechanics **15**, 934–950 (2021).
[16]W. Wang, F. Wang, D. Lai, and Q. Chen, "Evaluation of sars-cov-2 transmission and infection in airliner cabins," Indoor Air **32**, e12979 (2022).
[17]F. Wang, T. Zhang, R. You, and Q. Chen, "Evaluation of infection probability of Covid-19 in different types of airliner cabins," Build Environ **234**, 110159 (2023).
[18]T. Dbouk and D. Drikakis, "On coughing and airborne droplet transmission to humans," Physics of Fluids **32**, 053310 (2020).
[19]T. Dbouk and D. Drikakis, "On pollen and airborne virus transmission," Physics of Fluids **32**, in production (2021).
[20]T. Dbouk and D. Drikakis, "On respiratory droplets and face masks," Physics of Fluids **32**, 063303 (2020).
[21]T. Dbouk and D. Drikakis, "Weather impact on airborne coronavirus survival," Physics of Fluids **32**, 093312 (2020).
[22]T. Dbouk and D. Drikakis, "On airborne virus transmission in elevators and confined spaces," Physics of Fluids **33**, 011905 (2021).
[23]S. Bushwick, T. Lewis, and A. Montañez, "Evaluating COVID Risk on Planes, Trains and Automobiles," Scientific American (2020).
[24]V. Mathai, A. Das, J. Bailey, and K. Breuer, "Airflows inside passenger cars and implications for airborne disease transmission," Science Advances **7**, eabe0166 (2021).
[25]V. Mathai, A. Das, and K. Breuer, "Aerosol transmission in passenger car cabins: Effects of ventilation configuration and driving speed," Physics of Fluids **34**, 021904 (2022).

[26]K. Luo, Z. Lei, Z. Hai, S. Xiao, J. Rui, H. Yang, X. Jing, H. Wang, Z. Xie, P. Luo, W. Li, Q. Li, H. Tan, Z. Xu, Y. Yang, S. Hu, and T. Chen, "Transmission of SARS-CoV-2 in Public Transportation Vehicles: A Case Study in Hunan Province, China," Open Forum Infectious Diseases **7**, ofaa430 (2020).
[27]Y. Shen, C. Li, H. Dong, Z. Wang, L. Martinez, Z. Sun, A. Handel, Z. Chen, E. Chen, M. H. Ebell, F. Wang, B. Yi, H. Wang, X. Wang, A. Wang, B. Chen, Y. Qi, L. Liang, Y. Li, F. Ling, J. Chen, and G. Xu, "Community Outbreak Investigation of SARS-CoV-2 Transmission Among Bus Riders in Eastern China," JAMA Internal Medicine **180**, 1665–1671 (2020).
[28]P. Azimi, Z. Keshavarz, J. Laurent, B. R. Stephens, and J. G. Allen, "Mechanistic transmission modeling of covid-19 on the diamond princess cruise ship demonstrates the importance of aerosol transmission," medRxiv (2020), 10.1101/2020.07.13.20153049.
[29]L. Moriarty, M. Plucinski, and e. a. Marston, B.J., "Public health responses to covid-19 outbreaks on cruise ships — worldwide," MMWR Morb Mortal Wkly Rep **69**, 347–352 (2020).
[30]E. C. Rosca, C. Heneghan, E. A. Spencer, J. Brassey, A. Plüddemann, I. J. Onakpoya, D. Evans, J. M. Conly, and T. Jefferson, "Transmission of SARS-CoV-2 Associated with Cruise Ship Travel: A Systematic Review," Tropical medicine and infectious disease **7**, 290 (2022).
[31]L. Huang, I. Riyadi, S.and Utama, M. Li, P. Sun, and G. Thomas, "Covid-19 transmission inside a small passenger vessel: Risks and mitigation," Ocean Engineering **255**, 111486 (2022).
[32]A. Saunders, "Cruise lines change ship ventilation systems as part of overall covid strategy," (2020).
[33]K. Wiles, "Cruise ship ac systems could promote rapid coronavirus spread, prof says," (2020).
[34]O. Almilaji, "Air Recirculation Role in the Spread of COVID-19 Onboard the Diamond Princess Cruise Ship during a Quarantine Period," Aerosol and Air Quality Research **21**, 200495 (2021).
[35]J. Zhou, S. P. Chen, W. W. Shi, M. Kanrak, and J. Ge, "The impacts of COVID-19 on the cruise industry based on an empirical study in china," Marine policy **153**, 105631 (2023).
[36]M. Z. Bazant and J. W. M. Bush, "A guideline to limit indoor airborne transmission of COVID-19," Proc Natl Acad Sci USA **118** (2021), 10.1073/pnas.2018995118.
[37]CDC, "COVID-19 Ventilation in Buildings 2023," The Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/community/ventilation.html (2023).
[38]ASHRAE, "ANSI/ASHRAE Standard 241-2023, Control of Infectious Aerosols," (2023).
[39]ASHRAE, "ANSI/ASHRAE Standard 62.1-2019, Ventilation and Acceptable Indoor Air Quality," (2019).
[40]WHO, "Roadmap to improve and ensure good indoor ventilation in the context of COVID-19," World Health Organization. https://www.who.int/publications/i/item/9789240021280 (2021).
[41]REHVA, "COVID-19 guidance 4.1, how to operate HVAC and other building service systems to prevent the spread of the coronavirus (SARS-CoV-2) disease (COVID-19) in workplaces," Federation of European Heating, Ventilation and Air Conditioning Associations. https://www.rehva.eu/activities/covid-19-guidance/rehva-covid-19-guidance (2021).
[42]Y. Li, P. Cheng, and W. Jia, "Poor ventilation worsens short-range airborne transmission of respiratory infection," Indoor Air **32**, e12946 (2022).
[43]FPS, "Legal framework regarding indoor air quality," Federal Public Service - Public Health. https://www.health.belgium.be/en/closer-legal-framework-indoor-air-quality (2022).
[44]J. G. Allen and J. D. Macomber, *Healthy Buildings: How Indoor Spaces Can Make You Sick–Or Keep You Well*, 2nd ed. (Harvard University Press, 2022).
[45]ASHRAE, "ANSI/ASHRAE Standard 62.2-2019, Ventilation and Acceptable Indoor Air Quality in Residential Buildings," (2019).
[46]K. Ritos, D. Drikakis, and I. Kokkinakis, "Virus spreading in cruiser cabin," Physics of Fluids (2023), https://doi.org/10.1063/5.0169992.
[47]X. Zhao, S. Liu, Y. Yin, T. T. Zhang, and Q. Chen, "Airborne transmission of covid-19 virus in enclosed spaces: An overview of research methods," Indoor Air **32**, e13056 (2022).
[48]R. Dhand and J. Li, "Coughs and sneezes: Their role in transmission of respiratory viral infections, including sars-cov-2," AJRCCM **202** (2020), 10.1164/rccm.202004-1263PP.
[49]W. van der Reijden, E. Veerman, and A. Nieuw Amerongen, "Shear rate dependent viscoelastic behavior of human glandular salivas," Biorheology **30**, 141–152 (1993).
[50]W. E. Ranz and W. R. Marshall, "Evaporation from drops, Part I," I. Chem. Engng. Prog. **48**, 141–146 (1952).
[51]W. E. Ranz and W. R. Marshall, "Evaporation from drops, Part II," I. Chem. Engng. Prog. **48**, 173–180 (1952).
[52]Siemens Digital Industries Software, "Simcenter STAR-CCM+, version 2210," (2022).
[53]B. E. Scharfman, A. H. Techet, J. W. M. Bush, and L. Bourouiba, "Visualization of sneeze ejecta: steps of fluid fragmentation leading to respiratory droplets," Exp Fluids **57**, 1–9 (2016).
[54]S. Zhu, S. Kato, and J. H. Yang, "Study on transport characteristics of saliva droplets produced by coughing in a calm indoor environment," Building and Environment **41**, 1691–1702 (2006).
[55]X. Xie, Y. Li, H. Sun, and L. Liu, "Exhaled droplets due to talking and coughing," Journal Royal Society Interface **6**, 703–714 (2009).
[56]W. Weibull, "A statistical distribution function of wide applicability," Journal of Applied Mechanics **18**, 93–297 (1951).
[57]Y. Liu, Y. Laiguang, Y. Weinong, and L. Feng, "On the size distribution of cloud droplets," Atmospheric Research **35**, 201–216 (1995).
[58]J. Lee and G. Lee, "Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation," Neural Networks **161**, 682–692 (2023).
[59]J. Lee and G. Lee, "Unsupervised domain adaptation based on the predictive uncertainty of models," Neurocomputing **520**, 183–193 (2023).
[60]Z. Mousavi, T. Yousefi Rezaii, S. Sheykhivand, A. Farzamnia, and S. Razavi, "Deep convolutional neural network for classification of sleep stages from single-channel eeg signals," Journal of Neuroscience Methods **324**, 108312 (2019).
[61]Z. Mousavi, S. Varahram, M. Mohammad Ettefagh, and M. H. Sadeghi, "Dictionary learning-based damage detection under varying environmental conditions using only vibration responses of numerical model and real intact state: Verification on an experimental offshore jacket model," Mechanical Systems and Signal Processing **182**, 109567 (2023).
[62]R. P. Feynman, *The Character of Physical Law* (MIT Press, Cambridge, MA, 1965) iSBN: 978-0262560030.
[63]R. Ariew, *OCKHAM'S RAZOR: A HISTORICAL AND PHILOSOPHICAL ANALYSIS OF OCKHAM'S PRINCIPLE OF PARSIMONY.* (University of Illinois at Urbana-Champaign, 1976).
[64]B. Dresp-Langley, O. K. Ekseth, J. Fesl, S. Gohshi, M. Kurz, and H.-W. Sehring, "Occam's razor for big data? on detecting quality in large unstructured datasets," Applied Sciences **9**, 3065 (2019).
[65]F. J. Bargagli Stoffi, G. Cevolani, and G. Gnecco, "Simple models in complex worlds: Occam's razor and statistical learning theory," Minds and Machines **32**, 13–42 (2022).
[66]A. A. Neath and J. E. Cavanaugh, "The bayesian information criterion: background, derivation, and applications," Wiley Interdisciplinary Reviews: Computational Statistics **4**, 199–203 (2012).
[67]Y. Lu and Z. Lin, "Coughed droplet dispersion pattern in hospital ward under stratum ventilation," Building and Environment **208**, 108602 (2022).